DOCUMENT RESUME

ED 255 568 TM 850 211

AUTHOR Cliff, Norman
TITLE Ordinal Measurement.
PUB DATE 26 Aug 84
NOTE 30p.; Paper presented at the Annual Meeting of the American Psychological Association (92nd, Toronto, Ontario, Canada, August 24-28, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Multiple Regression Analysis; *Questionnaires; Research Problems; Scores; *Statistical Analysis; *Test Interpretation; Test Theory; *True Scores
IDENTIFIERS *Ordinal Scales

ABSTRACT

In almost all applications of measurement there is some sort of response by a human subject. Almost always, the response scale is ordinal, but almost always it is treated as if it were an interval measure. Methods for treating data ordinally are currently being developed in three areas: ordinal analysis for questionnaire responses, ordinal approaches to multiple regression, and an ordinal true score theory. Each of these areas is discussed in detail. In all of them, a major concern has been to stay as close as possible to the form in which the data are presented. When the data are dichotomous, statistics are used that measure the degree of agreement with the least distortion due to extraneous factors. Least squares statistics are avoided as a criterion of fit whenever possible. Kendall's tau, or one of its variations, is used as a measure of correlation. While success is not complete in these ordinal data analysis methods, there is enough of it to suggest that it could be extensive. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED255568

Ordinal Measurement

Norman Cliff

University of Southern California

Los Angeles, California 90089-1061

Division 5 Presidential Address, presented August 26, 1984, Toronto, Ontario, Canada

Please do not cite without permission

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

N. Cliff

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 852 211

Ordinal Measurement

The Problem of Ordinality

In almost all applications of measurement there is some sort of response by a human subject. Almost always, the response scale is ordinal, but almost always it is treated as if it were not.

What we are talking about here are rating scales, Likert questionnaire response scales, dichotomous or polytomous test items, and the like. Surely these are all ordinal measures, but equally surely they are treated as if they were interval measures by most of us in most applications. Most of us know about the difference between ordinal and interval scales, so why do we do it, is there any cost to doing it, and are there any alternatives?

We treat the data as if it were interval for one of two reasons, I think. One is that it is easier to do so, and the other is that we think we get more information from the data when it is treated as continuous interval variables. Clearly, it is easier to treat the numbers in a response scale as if they represented equal intervals than it is to treat them ordinally, because finding ranks is a rather laborious process. Furthermore, the computer programs most readily available to us may not furnish ordinal options. The usual parametric statistics, means, variances, correlations, and so on, seem to us to provide more information than the rather more awkward ordinal alternatives. Furthermore, when we apply powerful metric models such as

factor analysis or latent trait theory to the data, the results have a satisfying ring of authenticity.

It is also true that the assumption of continuity for the variables provides technical simplifications of developing statistical methodology. Continuous variables are very much easier to work with mathematically, so it is simpler to develop and fit models for continuous data. This is a major reason for the scarcity of technology for working with data ordinally.

But why should we worry? Why should we not follow the example of Mad Magazine's Alfred E. Newman? Most people in measurement are practical people, and so does it make any practical difference to treat data as interval when they are ordinal? Certainly there are no reported fatalities from the practice. Furthermore, a long time ago Abelson and Tukey (1963) showed that the correlation between a scale and any reasonable transformation of that scale would be pretty high. Nevertheless, I think that there are practical consequences; that fairly often our interval treatment of ordinal data leads us at least somewhat astray.

Practicality isn't everything, though. People in measurement also tend to have a strong sense of wanting to do the right thing, not just the easiest thing. We are uneasy about treating our data in a way that we worry is not justified. One of the most frequent questions that comes to a statistical consultant is the one of what to do about ordinal data. Even if she tells the student or colleague

to just ignore the problem for the reasons above, she may do so with some regret and unease about the possible validity of the conclusions that will be drawn. Besides, if our data are ordinal, they are telling us about orders, so it might just feel better to be consistent, and have our analyses and conclusions tell about orders also. Today I'll describe some attempts at developing methods for ordinal analyses of data.

Proposals

Over the past several years, some colleagues and students and I have been trying to improve the situation with respect to ordinal models, by developing methods for treating data ordinally, a kind of ordinal psychometrics. This paper sketches some of the more recent developments in this program.

The three things that will be described in a little detail are first an ordinal analysis for questionnaire response, second some ordinal approaches to multiple regression, and third an ordinal true score theory. Some other things we have done that I will not describe are an ordinal approach to observed item scores (Cliff, 1977, 1979), a cluster analysis for binary data (Cliff, et al, 1984), a set of axioms to characterize three-mode Guttman scales (Collins, 1983).

In all these methods, a major concern has been that we stay as close as possible to the form in which the data are presented. When the data are dichotomous we try to use statistics that measure

the degree of agreement with the least distortion due to extraneous factors. We try to avoid least squares as a criterion of fit whenever possible. We lean toward Kendall's tau in one or the other of its multiple forms as the measure of correlation. These are not rigid policies; as a matter of expedience, we do use the more conventional statistics and criteria on occasion. For the most part, though, we try to follow the ordinal bias I've indicated.

Ordinal Treatment of Questionnaire Response

Except for test responses with a right-wrong response format, the most common type of data is questionnaire responses with an agree-disagree format. Both furnish ordinal information, but they furnish it in very different ways.

Questionnaire items are typically treated in the same way as ability test items, but they represent a type of response that is fundamentally different from a measurement theory point of view. If we have a collection of statements presumably reflecting different degrees of say self-esteem these are quite different in their behavior from a set of mathematics items. In a mathematics item that says "solve $a + bx = c$ for x," we think of there being essentially only one kind of person who gets the item wrong: the person who has not learned to solve linear equations. But consider a self-esteem item such as "In most situations, I am the most capable person present." The

positive or keyed response is "true" or "agree," but there are two kinds of people who would get it "wrong," i.e., mark disagree. There are the obvious humble folk who feel that most of the time they are not the most capable person. Then there is also that extreme few who mark "disagree," snorting, "What do you mean, 'Most situations?'"

Thus the probability of a keyed response is curvilinear with the underlying trait. Some people call this kind of response "unfolding," although that terminology I find a little inaccurate.

These two kinds of positive responses can be called a dominance response in the case of the algebra item; that is, the person dominates the item, and a proximity response in the case of the questionnaire item, that is, the person is close to the item. Each kind of response can be used to order the persons who respond, provided there is a continuum of items, but the process by which we construct the order should be different, or so one would think.

If the persons and items are ordered correctly for a set of mathematics items, then the item response matrix will look like Figure 1a, the familiar triangle of a Guttman scale, ignoring minor problems like unreliability. On the other hand, the correct order for the proximity response ideally will look like Figure 1b, a parallelogram form (Coombs, 1964) that is less familiar to most of us, except for students of Clyde Coombs and recent students of my own.

Instert Figure 1 about here

The practice, of course, is to treat both kinds of items the same.

The items toward the right in diagram 1b presumably represent low self esteem, and the ones in the middle represent intermediate. Conventional item analysis procedures, such as factor analysis or correlating items with the total score on a selected cluster, or some combination of the two, will indeed yield a self esteem scale. The items toward the right correlate negatively with the high self esteem items, those in the middle will correlate low with them. This means that the right-hand will be keyed negatively on the self-esteem scale and the middle ones will be discarded. When those steps are taken and the items are appropriately re-ordered, the item response matrix as reduced and rekeyed will look like Figure 1c. Many of us have carried out such analyses many times, with reasonably satisfactory results.

But such analyses bother me. I like Figure 1b; it seems both neat and natural. The 1's in the matrix all mean "agrees with," whereas in 1c the 1's mean "answers in the direction that the investigator thinks is high self esteem," and as a general rule the investigators predilections should intrude in our data as little as possible.

Now response matrices do not come to us in the nice neat form of 1b. At best they are more or less arbitrarily re-ordered. For example, Figure 1d is a random reordering of 1b. Furthermore, there will be a certain amount of inconsistency, so that the optimum order may not be so readily apparent even if it is found. Therefore, if we think that 1b is the correct prototype rather than 1a, we need some means of finding

the optimum order for the rows and columns so that it most nearly brings things into this parallelogram form.

Minimizing Variances of Ranks

If there is no inconsistency in the data, then there are lots of ways to find the appropriate order. However, if there is inconsistency, then they run into difficulties. A principle that has become clear in the last 20 years is that most of what we call technical psychometrics is the process of fitting models to data through the optimization of some measure of the fit of the model. Such measures of fit are called objective functions, among other things, and common examples are least squares, maximum likelihood, minimizing stress, and the like. The criterion used in any one application has to be chosen carefully in the light of the properties of the data. What should be used here to define the optimum order? If we compare 1b to the scrambled version in 1d, the main difference is that the 1's in each row and in each column of 1b are bunched together. For the time being, let us concentrate on the rows and assume that the same reordering process can be applied to columns. We want to maximize the the extent to which the 1's in each row are bunched together. What is the most obvious measure of the bunchiness? Most of us will think of the variance as a measure of spread, and that is what was used in our work: the average variance of the ranks of the endorsed items in each row. If you would like to see a formula, here is one

$$\bar{s}^2 = \frac{1}{n} \sum_i \left[\sum_j (x_{ij} r_j - \bar{r}_{i.})^2 / \sum_j x_{ij} \right]$$

In the formula, x_{ij} is the dichotomous response variable, r_j is the rank order of item j, and n is the number of rows or respondents. We want to choose the ranks r_j in such a way as to minimize this function, and then reorder the columns accordingly.

Now the question is how to do it. If the r_j can take on any values, that is if the item scale is allowed to be continuous, then the problem is an old one. This is optimum scaling, or what the French have elegantly rediscovered as correspondence analysis (Tenenhaus & Young, in press). But we are restricting ourselves to ranks for the r_j, so the conveniences or principal components as a way of finding the r_j are denied to us. We have to fall back on finite methods.

The main operation in the methods we have developed in the pairwise interchange of columns so as to reduce the objective function as much as possible. This is an iterative process. At iteration 0, the beginning, we have an arbitrary set of ranks for the r_j, yielding a value for $\bar{s}^2$; then we find the implied effect on $\bar{S}^2$ of interchanging each pair of columns and interchange the pair that will reduce it the most. Then we find the variance again and the potential effect of interchanging each pair, interchange the one having the largest effect, and so on, continuing the process until no pairwise interchange will result in an improvement.

It would be nice if that were the end, and sometimes it is, but it is not always, for it is not uncommon to reach a cul-de-sac, the analogue of a local minimum in a continuous optimization problem. That is, no pairwise interchange improves matters, but we have not really reached the minimum for $\bar{s}^2$. We have therefore introduced a couple of

additional procedures. One searches for a particular kind of distortion that we found to be pretty common in some data we have analyzed. The other adjustment essentially gives the supposed final order a series of random kicks followed by further pairwise interchanges to see if the same solution or a better one is found. Finally, a different random starting order is generated, and the whole process starts again from the beginning. In this way, several different solutions may be generated, or the process may arrive at the same place several times. If there are different solutions, the investigator chooses the best one, presumably by the criterion of smallest variance of ranks, but perhaps tempered by some knowledge of the data.

There is apparently no absolute way of guaranteeing arrival at the best possible solution for this problem. It is n-p complete, a variant of the infamous traveling salesman problem, and so there is no guarantee that the solution found is the best one unless we have tried all possible orders, an impractical approach with more than a few items since if p = 10, p! is 3,628,800 and for p = 20 is roughly 2×10^{18}.

This uncertainty about the solution is a price we pay for moving from a continuous model to an ordinal one. I do not mind paying it for two reasons. One is that I would rather work with the ordinal model than with a continuous one. The other is that using a continuous model, such as optimum scaling, may well give back a solution that is not very good by our criterion, suggesting that if minimizing the average variance

of ranks is even more or less what we have in mind, then the methods that assume a continuous interval scale may not give very valid results.

Our approach has been tried out on several sets of data. Some of the results are illustrated in Figures 2, and 3. For the illustrations, we use a box or blank instead of 1 and 0. The three sets in Figure 2 are questionnaire response.

It should be understood that in these applications the pairwise interchange process and associated adjustments have to be applied twice, once to the rows and once to the columns. This can be rather time-consuming for the long dimension of the matrix, usually representing persons, so we have an approximate procedure to apply to the long dimension, once the short dimension has been ordered. This is to order the rows of the matrix in terms of the median rank of the endorsed items for that row. If there is anyone who still remembers Thurstone attitude scaling (Torgerson, 1958), this will be reminiscent of the way Thurstone attitude scales were applied.

Before leaving this method, let me point out that it can be applied in a wide variety of contexts. We have emphasized instances where the response matrix is persons by items, but is would also apply in a variety of other cases where a single underlying continuum is expected. The last data, Figure 3, illustrates the bredth of application of such a method. It is archaeological data, a matrix of grave sites by artifact types, where a positive entry means that the artifact type was found in that grave.

The order from upper left to lower right is presumed to represent chronological order of burial. The matrix could also be legislators by bills, and the response would be "voted for." Or it could be stimulus by stimulus, and the response is "is similar to." In the latter case, we order the rows and columns simultaneously instead of separately. The latter trick can also be used if we have a multilevel rather than simply a dichotomous response. In that case, with k levels of response, we have k - 1 matrices, one for each possible dichotomy: highest versus the rest, highest or next highest versus the rest, and so on. This has the effect of weighting the higher levels of endorsement higher.

Finally, let me point out that minimizing the ranks of the endorsed items within each row is a reasonable criterion in the case of the triangular response matrix also, i.e., dominance data, as our method is likely to give the right order, whether 1a or 1b is the prototype.

The point of this example has been to show that it is feasible to treat ordinal questionnaire response data with ordinal methods. The ordinal methods are feasible and reasonably effective. There are limits to the size of problem they can be applied to in their present form, but we have made little effort to make our programs efficient, so considerably better efficiency could be expected. A bit more detail on this method is given by Cliff (1984) and by Zatkin, et al.,(1983).

Ordinal Multiple Regression

In many applications statistical prediction is based on what are

really ordinal variables. Suppose we have ordinal predictors and an ordinal criterion and would like to combine the orders on the predictors so as to optimally predict the order of the criterion. What do we do? The conventional approach is to ignore the ordinal nature of the data and do multiple regression, minimizing the sum of squares of deviations on the criterion scale as it is given. This seems arbitrary and unrealistic. An alternative is to convert all the data to ranks and then do the multiple regression. This latter seems satisfactory to some people, but to me the ranks are just one of the many allowable transformations of the ordinal scale and minimizing the sum of squared deviations on that form is not satisfactory either. It would be preferable to stay closer to the ordinal nature of the data than either of these do.

Notice that the problem was stated as one of combining orders to predict orders. For example, we may wish to use several four point rating scales from an interview to predict on-the-job performance or to predict post-hospitalization adjustment, again measured on a four point scale, or some other ordinal measure. In each case, we would like to treat the scales only ordinally.

Now it is necessary to think about things in a new way, because an ordinal relation refers to a pair of entities, not to a single entity. What we are saying is that we want to use the order relations of each pair of persons on the predictors to predict their order relations on the criterion.

What follows here owes a good deal to suggestions by Reynolds and Sutrick, but many of the details are new. An order relation can be treated in a numerical way. For every pair, h and i, says, there is a number, d_{hi}, the dominance of h over i where

$d_{hi} = 1$ if h ranks ahead of i;

$d_{hi} = -1$ if h ranks behind i;

$d_{hi} = 0$ if h and i are tied.

We can see that $d_{hi} = -d_{ih}$ so we only need to record one of them. That is, we take the $\frac{1}{2}n(n-1)$ possible pairs of persons, and for each define the d_{hi} on each variable, so really we need subscript for variables. Letting y stand for the criterion we have d_{hiy} as the ternary variable that tells us the order relation for pair h, i on the criterion y. For our convenience, we can put the subjects in the same order as their criterion scores so that $h < i$ if $y_h \leq y_i$. This means that none of the d_{hiy} are negative, and this simplifies matters because now the sign of the dominances for the other variables, d_{hij} for predictor j, indicates whether or not the relation for that pair on that variable is the same as on the criterion. This is illustrated in Table 1. What we want to do it somehow "predict" all those positive d_{hiy} from the mixed positive and negative d_{hij}.

There are lots of ways to do it, depending on how we structure the question. A straightforward way is to say that we want to minimize

$$\phi_1 = \sum_{n,i} (d_{hiy} - \hat{d}_{hiy})^2$$

where $\hat{d}_{hiy} = \sum_j w_j d_{hij}$, for some weights w_j.

That is, we can make a least-squares problem out of it. This has a straightforward and well-known solution, the same one here as in multiple regression, or wherever we want to find a set of weights to make a least-squares prediction:

$$w_j = (D'D)^{-1} D' \underset{\sim}{d}_y$$

In fact, we could put the dominance matrix in the computer, and do m.r., but it would have to be m.r. with zero intercept. Here, D is the matrix of dominance variables on the predictors, and $\underset{\sim}{d}_y$ is the vector of dominance scores on the criterion. Table 1 goes on to show the weights for these data and the "predicted" dominance relations on the criterion. Rather interestingly, this solution for the w's can be expressed in terms of Kendall's tau rank correlations:

$$w_j = T^{-1} \underset{\sim}{\tau}_y .$$

Here, T is a matrix of the Kendall taus among the predictors, with the proportion of untied pairs in the diagonal, and $\underset{\sim}{\tau}_y$ is the vector of taus of the predictors with the criterion. This is interesting, because of the parallel with ordinary multiple regression, except it is telling us to use the tau correlations rather than the Pearsons.

If we do not like this quasi-quantitative approach to ordinal multiple

regression, are there any alternatives? Let me mention a couple. We have arranged things so that all the dominance relations on the criterion are positive unities, so we can make our predictions more qualitative, more like them, by recording the d_{jiy} so that if they are positive, we call that a correct prediction and code it +1 if they are negative, we call it incorrect and code it -1. If any happen to be exactly zero, that is neutral. Now the sum of products of predictions and criterion is a tau. Now, by the criterion, the "multiple regression" problem is to choose the weights to apply to the preditors so that as many as possible of the $\hat{d}$s are positive. This is a very ill-defined problem, because we could fiddle with the weights indefinitely and never know whether we had arrived at the best solution, and it becomes more and more difficult the more predictors we have.

Rather paradoxically, we can make this problem easier by making it as hard as possible. Suppose there are a whole lot of predictors and a whole lot of people and we apply any arbitrary set of weights. The statisticians' friend, the central limit theorem, comes racing to our rescue because the distribution of the $\hat{d}$s will be normal. If the positive $\hat{d}$s are called correct, and the negative ones incorrect, then any given vector of weights will give us some of each. The proportion of positive and negative $\hat{d}$s just depends on how far the mean of the $\hat{d}$s is from zero in standard deviation units. This tells us that what we want to use for weights is the weights that maximize the ratio of the mean $\hat{d}$ to the standard deviation of the $\hat{d}$s. Maximizing this kind of ratio by means

of weights is a very common problem in statistics, and we can find the weights that do this. The result was a little surprising, because it turns out, via an algebraic coincidence, that the weights that we got before, for minimizing the sum of squared deviations, will work for this purpose also.

We can find the proportion of these "predictions" that are qualitatively correct, because it is the proportion of this normal curve that falls above zero. We can find that proportion by finding the mean and standard deviation of the predictions and from them the z-score for zero and from it the corresponding proportion. Kendall's tau is the proportion of pairs were the order is the same minus the pairs where they are different, so subtracting one minus the proportion from the proportion will give us a "multiple tau" defined in this way.

Note that in order to justify this, we had to assume that there were a lot of predictors and subjects, but the approximation may be reasonable even if there is only a few of each. With our little example, it turned out that eight of the ten predictions were positive, .8 minus .2 is .6, so we have a "multiple tau" here of .6. Now we can compare this to the process that assumed normality of the $\hat{d}$s. The mean $\hat{d}$ is .4036 and the s.d. is .4906, giving a z-score for zero of -.823. The proportion of the normal curve that falls above -.823 is .795. .795 minus its complement is .590, remarkably close to the actual .60 even though the $\hat{d}$s in this little example are not very normally distributed. This correspondence is perhaps a coincidence, but if so it si an encouraging one.

We might like to take an additional step in making our prediction process even more qualitative or ordinal. This is to limit ourselves to ternary weights, weights for the predictor dominance that are themselves +1, -1 or 0. +1 means dominances on that predictor are counted , -1 means that they are to be reserved, and 0 means we ignore that variable. As before the overall predicted dominance will be counted as correct if the weighted sum of predictor dominances is positive, incorrect if it is negative, and neutral if it is zero. So, for example, if we use weights of 1, -1, 1, the prediction for the first pair in the table is -3, for the second pair is -1, and so on down. Each of these positive totals is a correct prediction of order. The total for the 2,3 pair would be -1, so this is incorrect. What we want to do again is to choose the weights from this restricted set in such a way as to make as many positive predictions as possible.

There is no way to do this analytically, and there are 3^p possible vectors of weights. What we can do is try some heuristics. For example, we could find the continuous weights by the method above and just use their signs. Second, having done that, we could explore the effect of changing the weights one at a time from the current value, reversing one or setting it to zero or from zero to +1 or to -1. If any such change results in an increase in accuracy, we make the change and go through the process again, until we find no change increases prediction. This

process is analogous to forward and backward selection multiple regression.

It is interesting that here the optimum weight for a variable will often be zero; setting a weight to zero from plus or minus one can well result in improvement in ordinal prediction. The most obvious case is where there are two predictors and no ties on the predictors. Then, it is always true that the best prediction is always done by just considering the better predictor by itself; i.e., weighting the poorer one zero. You can demonstrate that for yourself with our little data set here.

I think that the results here are interesting. We can do multiple regression of a sort with the ordinal properties of our data. We can think of several different ways to do it, based on several different philosophies. It turns out that two of them lead to the same procedure, and the procedure is determinate. The other leads to an iterative or exploratory process, but there seem to be feasible ways of carrying it out also.

Ordinal Test Theory

The third kind of application I want to talk about is ordinal test theory. To me, test items are the most obvious case where we habitually take ordinal data and treat them as if they were interval, or derive from them continuous scale scores through the assumption of one of the other models. These models are typically the latent trait or item response theory models that assume various kinds of tracelines

relating the ordinal response to an underlying continuum. It may be heresay to say it, but I am not comfortable doing that. I will exercise executive privilege by making some assertions about such processes without backing them up with data. I would like to have some substitute for item response theory because (a) it takes an enormous amount of data and facilities to apply it, and (b) its assumptions are almost certainly false for most tests. This is not to say that what I will tell you about is sufficient to replace this marvelous edifice that has been constructed in the past 20 years or so. It may, however, be sufficient to suggest the possibility that we can replace that edifice with a much simpler structure that provides just as much shelter.

A number of years ago, I suggested some ordinal ways of looking at test responses and of measuring their consistency (Cliff, 1977, 1979). This was at the level of the data itself, and the paper was called "Test theory without true scores?" Now I want to bring the idea of true scores back, but assume that they are ordinal. I prefer, by the way, to think of this true order as an order on a universe of items rather than as an undefined latent trait, but that is really not essential for what I am going to say here.

What we would like to do is define the ordinal validity or fidelity of a test, the extent to which a score from the test produces an order for the subjects that agrees with the universe order. We would measure such fidelity with the Kendall's tau between observed and true order. How can this be done?

I am not sure about the complete answer, but we seem to have parts of it. A couple of years ago, (Cliff, 1982) I tried to formulate an ordinal analogue of the local independence assumption that underlies item response theory. This turned out to be possible, although a little more complicated than just saying "local independence." It led to an expression for the ordinal fidelity of a test item that was a simple function of the ordinal relation between two test items. It also gave a similarly simple expression for the fidelity of the total score on a test. Unfortunately, these results were limited to two-item tests, The market for two-item tests is rather low, so matters stayed there for a while.

The results concerning multiple regression that are described above appear to offer an out, and that is a direction currently being followed. In a test, we are combining dominances on a number of items to yield an overall score that is an estimate of the order on the criterion. In this case we are using item orders to predict score order. In the ordinal case, as we have seen, the relevant scores are these three-valued variables for pairs of persons. Weights are applied to the predictors to estimate the criterion. What we need in order to find and evaluate the weights and the correlation, is the matrix of taus among the predictors, and their taus with the criterion. We have the former, but not the latter since these represent here the unknown fidelities of the items. It seems that it should be possible to estimate these fidelities. At the very least, if the number of items is large, item-total

taus would estimate these fidelities. Alternatively, there are other assumptions that would allow these to be estimated from the inter-item taus.

Let me describe one approach. We have a total test order that combines the dominances on several items. This means that the weights for the dominances that are used are all unity by definition. If the summed dominances across items for a pair is positive, that pair is correctly ordered. As before in multiple regression, the probability that a pair is correctly ordered, i.e., the tau correlation with true order, just depends on the ratio of the mean predicted dominance to the standard deviation.

Here, in contrast to multiple regression, none of these criterion dominances are known. This turns out not to matter: we can estimate their mean and sigma anyway, by combining our approach to multiple regression with the results for a two-item test in assuming local ordinal indepencence that were alluded to above. Actually, we make an assumption that is more plausible than assuming local independence for all pairs of items. Instead, it is only necessary that average local uncorrelatedness holds across pairs of items. This gives us a way to estimate the average ordinal fidelity of the items. If variation in item fidelity is small relative to this average, or if it is uncorrelated with the item variances, then the average item fidelity is sufficient to estimate the overall fidelity of the test. That is, by this series of approximations we

can estimate the rank-order correlation between total score on the test and the underlying true order.

We can thus apparently develop a test theory that stays within the ordinal bounds of the data. It also seems likely that the solutions will be computationally simpler than the ones that continuous true score theories require and that they may be formulated with assumptions about item pools that are less unrealistic. Some of this lies in the future, but much of it seems very close to closure.

What I have tried to show overall here is some ways in which it has seemed possible to develop ordinal methods for treating our ordinal data. There was the case of parallelogram scaling, applicable, I think, to many kinds of self-report data. There are ordinal analogues of multiple regression that seem to provide many of the benefits of that technique. While success is not complete in these efforts, there is enough of it to suggest that it could be extensive. Perhaps the largest obstacle is our habits of thought that make it difficutl to think of these things.

References

Abelson, R.P., & Tukey, J.W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order, Annals of Mathematical Statistics, 34, 1347-1369.

Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 43, 375-399.

Cliff, N. (1979). Test theory without true scores? Psychometrika, 44, 373-393.

Cliff, N. (1984). An ordinal approach to parallelogram scaling. Paper presented at the June, 1984 meeting of the Psychometric Society, Santa Barbara, CA.

Cliff, N., McCormick, D.J., Zatkin, J., and Collins, L. Binclus: Nonhierarchical clustering of binary data. Unpublished manuscript, 1984.

Collins, L.M. (1983). The longitudinal Guttman simplex: a new scaling model for longitudinal data. University of Southern California, Los Angeles. Unpublished Ph.D. dissertation.

Coombs, C.H. (1964). A theory of data. New York: Wiley.

Reynolds, T.J. & Sutrick, K.H. (1985) Assessing the correspondence of a vector to a symmetric matrix using ordinal regression. Psychometrika, 50. In press.

Tenenhaus, M. & Young, F.H. (1985). Quantifying categorical multivariate data: An analysis and synthesis of many approaches. Psychometrika, In press.

Torgerson, W.S. (1958). Theory and methods of a scaling. New York: Wiley.

Zatkin, J., Cliff, N., Collins, L.M., & McCormick, D.J. (1983). Parallelogram scaling of binary items. Paper presented at the August 1983 meeting of the American Psychological Association, Anaheim, CA.

TABLE 1

STEPS IN ORDINAL REGRESSION

OBSERVED RANKS

PERSON	X_1	X_2	X_3	Y
1	1.5	4	1	1
2	4.5	1	3.5	2
3	1.5	4	[illegible]	3
4	3	2	2	4
5	4.5	4	3.5	5

DOMINANCES

PAIR	D: D_{X_1}	D: D_{X_2}	D: D_{X_3}	D_Y	T	τ_Y	$\hat{D}_Y$	$\hat{D}_Y^*$
1,2	+1	-1	+1	+1	.8 -.4 .2	.4	.343	+1
1,3	0	0	+1	+1	-.4 .7 0	.1	.169	+1
1,4	+1	-1	+1	+1	.2 0 .9	.3	.343	+1
1,5	+1	0	+1	+1			.910	+1
2,3	-1	+1	+1	+1	W		-.006	-1
2,4	-1	+1	-1	+1	D_{X_1} .246		-.343	-1
2,5	0	+1	0	+1	D_{X_2} .188		.566	+1
3,4	+1	-1	-1	+1			.006	+1
3,5	+1	0	-1	+1	D_{X_3} .056		.572	+1
4,5	+1	+1	+1	+1			1.476	+1
T	.4	.1	.3					.6

	Items					
Persons	A	B	C	D	E	F
1	1	1	1	1	1	1
2	0	1	1	1	1	1
3	0	0	1	1	1	1
4	0	0	0	1	1	1
5	0	0	0	0	1	1
6	0	0	0	0	0	1

Guttman response pattern

(a)

	Items					
Persons	A	B	C	D	E	F
1	1	1	0	0	0	0
2	1	1	1	0	0	0
3	0	1	1	1	0	0
4	0	0	1	1	1	0
5	0	0	0	1	1	1
6	0	0	0	0	1	1

Parallelogram response pattern

(b)

	Items			
Persons	A	B	-E	-F
1	1	1	1	1
2	1	1	1	1
3	0	1	1	1
4	0	0	0	1
5	0	0	1	0
6	0	0	0	0

Rescored parallelogram response pattern

(c)

	Items					
Persons	C	E	B	A	F	D
4	1	1	0	0	0	1
1	0	0	1	1	0	0
5	0	1	0	0	1	1
3	1	0	1	0	0	1
6	0	1	0	0	1	0
2	1	0	1	1	0	0

Scrambled parallelogram response pattern

(d)

Figure 1. Response patterns.

Figure 2. Paralellogram ordering of questionnaires: upper is Self-esteem; middle is Attitude toward War; lower is Tolerance for Ambiguity.

Figure 3. Ordered archeological data

Data from Hodson, Kendall, & Tautu, p.242